PROTEINS: Structure, Function, and Genetics 31:247–257 (1998)

Protein Folding Simulation With Genetic Algorithm and Supersecondary Structure Constraints

Yan Cui,[1] Run Sheng Chen,[1]* and Wing Hung Wong[2]*
[1]*Laboratory of Protein Engineering, Institute of Biophysics, The Chinese Academy of Sciences, Beijing, China*
[2]*Program in Statistics, University of California, Los Angeles*

ABSTRACT **We describe an algorithm to compute native structures of proteins from their primary sequences. The novel aspects of this method are: 1) The hydrophobic potential was set to be proportional to the nonpolar solvent accessible surface. To make computation feasible, we developed a new algorithm to compute the solvent accessible surface areas rapidly. 2) The supersecondary structures of each protein were predicted and used as restraints during the conformation searching processes. This algorithm was applied to five proteins. The overall fold of these proteins can be computed from their sequences, with deviations from crystal structures of 1.48–4.48 Å for C_α atoms. Proteins 31:247–257, 1998.**

Key words: protein structure prediction; supersecondary structure; genetic algorithm; solvent accessible surface area; hydrophobic potential

INTRODUCTION

The success of protein structure prediction depends on our knowledge of protein structure, interactions, and folding mechanism. At present, there are only a few reliable conclusions: 1) Native structures of proteins are compact and have well-packed cores which are highly enriched in hydrophobic residues.[1,2] 2) Hydrophobic interaction is the driving force for protein folding.[3,4] Native structures of proteins have minimal solvent-exposed nonpolar surface areas.[5] 3) Globular proteins are organized as a structural hierarchy,[6,7] i.e., secondary structure, supersecondary structure, tertiary structure, and quaternary structure. 4) The proteins employ folding pathways to avoid extensively searching the whole conformation space. They fold by hierarchic condensation.[7] The folding pathway is suggested to be secondary structures, supersecondary structure, domains, and ultimately whole protein monomers.[8–10]

We developed a protein structure prediction algorithm based on this crude knowledge. First, the supersecondary structures were predicted with an artificial neural network method.[11] Then we searched for low-energy structures in the conformation space under the constraints suggested by the supersecondary structures. The energy function is very simple and has only two terms—a hydrophobic interaction term, and a van der Waals interaction term. We used a genetic algorithm to search the conformation space. According to our design of this model, hydrophobic interactions drive the peptide chain to fold; van der Waals forces are used to reject the incorrect compact structures during the hydrophobic collapse. Only the structures in which there is minimal conflict between the hydrophobic interactions and the van der Waals interactions can survive and become dominant during the competition and selection process of the genetic algorithm. This algorithm was applied to five proteins. The overall fold of these proteins were computed from their sequences by this algorithm, with deviations from crystal structures of 1.48–4.48 Å for C_α atoms (Table I).

There have been several important advances in computer algorithms intended to predict native 3-dimensional structures of globular proteins from their amino acid sequences using simple energy functions.[12–14] The novel aspects of our method are: 1) We set the hydrophobic potential to be proportional to the nonpolar solvent-accessible surface area (NSASA). Although this is a reasonable way of including the hydrophobic effects, it was not used in previous protein structure prediction algorithms.[12–14] The computation of solvent-accessible surface area was very time-consuming. It would be prohibitively slow to incorporate it into protein structure prediction schemes which need to sample a large number of structures. We developed a new algorithm to calculate the solvent-accessible surface area rapidly. With this method we can compute the NSASA of every sampled structure in an acceptable time. 2) The supersecondary structures were predicted and used to derive soft constraints for the conformation search process. The identification of such structures repre-

Contract grant sponsor: Chinese National Scientific Foundation; Contract grant number: 39392900; Contract grant sponsors: UCLA, the Institute of Mathematical Sciences at the Chinese University of Hong Kong.
*Correspondence to: Run Sheng Chen, Institute of Biophysics, The Chinese Academy of Sciences, Beijing 100101, P.R. China; or Wing Hung Wong, Program in Statistics, UCLA, 8142 Math Sciences Building, Los Angeles, CA 90095.
Received 1 August 1997; Accepted 13 November 1997

TABLE I. Summary of the Computed Proteins

Protein	E_{hp}	E_{vdw}	E_{hb}	E_{total}	DME (Å)
1ROP					
Crystal	79.0	−30.3	0.0	48.7	0.0
Ga	76.9	−31.7	0.0	45.2	1.48
1UTG					
Crystal	104.5	−28.3	0.0	76.2	0.0
Ga	88.6	−37.1	0.0	51.5	3.47
1CRN*					
Crystal	63.2	6.8	−15.0	41.4	0.0
Ga	64.6	25.6	−40.0	50.2	2.73
1R69					
Crystal	64.2	−43.1	0.0	21.1	0.0
Ga	59.3	−20.3	0.0	39.0	4.48
1CTF					
Crystal	69.1	−24.1	−45.0	0.0	0.0
Ga	84.6	41.4	−70.0	56.0	4.00

*The native disulphide bond constraints were not used in the simulation.

sents important progress along the folding pathway. The conformation space is greatly reduced with these constraints.

METHODS

Review of Supersecondary Structure Prediction

Supersecondary structure is defined as the combination of two secondary structural elements with a short connecting peptide between one to five residues in length. A short connecting peptide can have a large number of conformations. They play an important role in defining protein structures. A connecting peptide usually changes the trend of the protein backbones so as to form an antiparallel turn, a vertical corner, a twist, or just a slight bend in a peptide chain.[11] The conformations of the residues in the short connecting peptides are classified into five major types, namely, a, b, e, l, or t,[16] each represented by a region on the ϕ-ψ map, respectively (see Fig. 1b in Sun and Jiang[16]). Supersecondary structures are classified according to their component secondary structural elements, the length of the connecting peptide, and the type of residues in the connecting peptide. In a survey of 240 proteins,[16] it was found that there are 34 types of supersecondary structures which occur more than 5 times. Of these 34 types there are 11 types of supersecondary structures which occur more than 25 times. These 11 types of supersecondary structures are called frequently occurring supersecondary structures. The 34 types of supersecondary structures occurred altogether 766 times, among which the 11 frequently occurring supersecondary structures occurred 568 times. This result shows that about 75% of the short-connecting peptides which occurred more than 5 times belong to the 11 types of frequently occurring supersecondary structures. Sun et al.[11] developed an artificial neural network method to predict the 11 frequently occurring supersecondary structure: H-b-H, H-t-H, H-bb-H, H-ll-E, E-aa-E, E-ea-E, H-lbb-H, H-lba-E, E-aal-E, E-aaal-E, and H-l-E, where "H" and "E" represent α-helix and β-strand, respectively. Each of these corresponds to a well-defined 3-dimensional motif (see Fig. 4 in Sun and Jiang[16]). The method of Sun et al. was used for supersecondary structure prediction. The predicted supersecondary structure will not be rigidly imposed on the conformation. Rather, it will serve to define suitable constraints (Table II) on affected torsion angles. In this way, the size of the conformation space is greatly reduced. However, under these constraints the conformation is still highly flexible and the structure can take on various shapes that are vastly different from the native shape.

Peptide Chain Representation

Amino acids are represented at the united-atom level. Bond lengths and bond angles are always fixed at their ideal values (according to Biosym's residue library). All the peptide bond dihedral angles are fixed in the trans ($\omega = 180°$) conformation. The degrees of freedom in this reduced representation are the backbone and sidechain torsion angles ϕ, ψ, and χ (some residues have more than one sidechain torsion angle).

Potential Energy Function

Our potential function has two terms: a hydrophobic interaction and a van der Waals interaction term,

$$E_{total} = E_{HH} + E_{vdw}.$$

We define polar and nonpolar united atoms by their heavy atoms: carbon and sulphur are nonpolar; nitrogen and oxygen are polar. The hydrophobic potential is proportional to the solvent-accessible surface of nonpolar atoms,

$$E_{HH} = C_h \cdot NSASA$$

where C_h is a constant and is set to 0.031 and $NSASA$ (in units of Å^2) is the nonpolar solvent accessible surface area.

We use a cut-off of 8 Å for van der Waals interactions,

$$E_{vdw} = C_V \cdot \sum f_{vdw}\left(\frac{r_{ij}}{R_i + R_j}\right)$$

where C_v is a constant that is set to 0.1, r_{ij} is the distance between atom i and atom j, R_i and R_j are the van der Waals radii of atom i and atom j, and the summation is over all pairs of atoms with $r_{ij} < 8$Å. The function f_{vdw} is a van der Waals potential with a tapering-off at short distances (Fig. 3):

$$f_{vdw}(r) = \begin{cases} \dfrac{1}{r^{12}} - \dfrac{2}{r^6} & (r > 0.8) \\ C & (r \le 0.8). \end{cases}$$

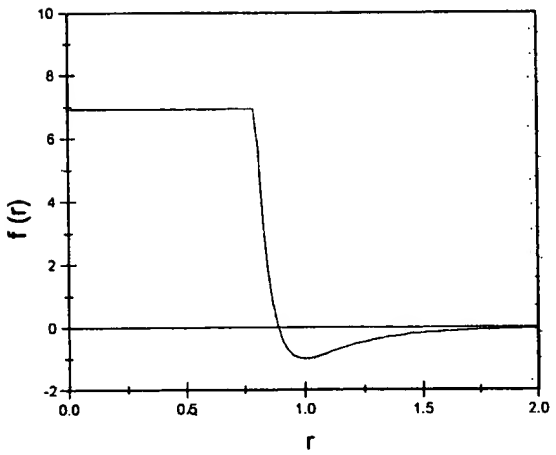

Fig. 1. The modified van der Waals potential energy function.

This tapered van der Waals potential will not reject a structure with low hydrophobic energy for only a few steric conflicts.

A backbone-to-backbone hydrogen bonding term is added for β-sheet structures. The distance between N···O pair should be no more than 3.5 Å and the out-of-plane dihedral angle between the oxygen and the peptide plane of the nitrogen ($C-N-C_\alpha$) should not exceed 40°. If these criteria are satisfied, an H-bond energy of −5.0 units is assigned.

Two-Level Lattice Method for Solvent-Accessible Surface Area Calculation

The hydrophobic effect is considered a principal force in the formation of the native protein structures.[3,4] A reasonable way of including the hydrophobic effect is to set the hydrophobic potential as proportional to the solvent-accessible surface area (SASA) of nonpolar atoms. Many algorithms for calculation of the SASA and molecule surface have been developed during the past 25 years.[17-40] Several fast numerical algorithms have been described in recent years.[38,39] However, calculation of the SASA has not been used in the folding simulation of whole protein molecules, in which a large number of SASA need to be calculated. In this work, we used a new algorithm to calculate the nonpolar solvent-accessible surface areas of about 300,000 conformations for each protein. Our method is based on the cube algorithm.[23-27] The whole molecule was put in a rectangular box. We introduced two levels of cubic lattice in the box, one was the coarse lattice (edge length 0.5 Å), the other was the fine lattice (edge length 0.1 Å), by which each cube in the coarse lattice was divided into 125 sub-cubes. (The cube in the fine lattice is called a "sub-cube," which differs from the cube in the coarse lattice.) Each of the 125 sub-cubes was assigned a number from 0 to 124. Water molecules were imitated by balls with a radius of 1.4 Å. Before calculating the SASA, we built a library of the cubic decomposition of the accessible surfaces and inner parts of each kind of atoms. We selected the center of a particular cube to be the origin. Then we put a "hydrated" sphere with a radius equal to the sum of the van der Waals radii of the atom and a water molecule around the atom at the center of one of the sub-cubes in this cube. If the center of a cube was covered by the sphere, it was marked by "V" (V-cube). Then every V-cube was checked. If a V-cube was on the surface, this meant that at least one of its six neighboring cubes was not a V-cube, then it was marked by "S" (S-cube). If a V-cube was not on the surface, it was marked by "I" (I-cube). Thus, we have two kinds of V-cubes, i.e., S-cubes, which were on the surface of the "hydrated" sphere, and I-cubes, which were at the inner part of the "hydrated" sphere. The totality of S-cubes was an approximation of the surface of the "hydrated" sphere. The totality of I-cubes was the cubic decomposition of the inner part of the "hydrated" sphere. The positions (lattice coordinates relative to the origin) of the S-cubes and I-cubes were recorded in the library. In such a way, the cubic decomposition of the "hydrated" sphere (include the surface and the inner part) whose center was at each of the 125 sub-cubes in the selected cube was recorded in the library according to the order of the sub-cube number.

With this library, the SASA can be calculated rapidly:

a) A protein molecule was put into the two-level lattice. For each atom, we determined which cube and sub-cube it was in. In other words, the Cartesian coordinates were transferred to lattice coordinates.
b) If an atom was in the cube at (lx, ly, lz), which were the lattice coordinates of the center of the cube, and in the sub-cube whose number is n, then we look for the record of the cubic decomposition of the "hydrated" sphere surface of this atom. The record was for the cube at the origin, so we translated them to (lx, ly, lz). In such a way, we put the cubic approximation of the "hydrated" sphere surface of every atom in the lattice. The cubes that were occupied by the "hydrated" sphere surface were marked by "S."
c) In the same way, we put the cubic approximation of the inner part of each atom on the lattice. These cubes were marked by "I." So, the S-cubes which were covered by the inner part of other "hydrated" spheres were remarked "I."
d) Every S-cube was checked to ensure that it was really on the surface of the "hydrated" protein molecule.
e) We counted the number of S-cubes. The number was proportional to SASA. Similarly, the total number of S-cubes belonging to nonpolar atoms would be proportional to NSASA.

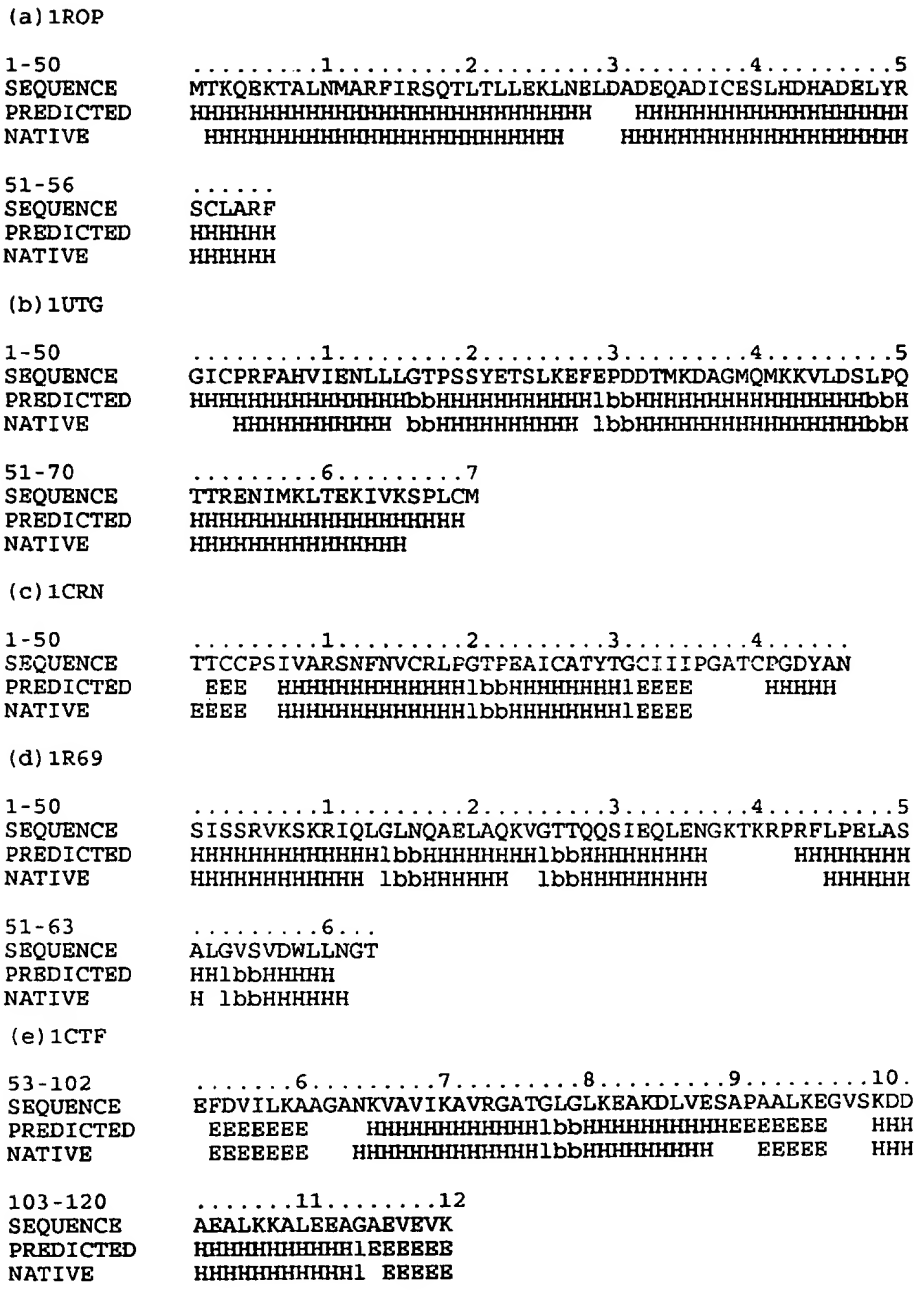

Fig. 2. Comparison of the predicted supersecondary structures (predicted) and the X-ray-elucidated supersecondary structure (native) of (a) 1ROP, (b) 1UTG, (c) 1CRN, (d) 1R69, and (e) 1CTF.

The advantage of this method is that it does not require any calculation of distances between points. Besides table look-up, it only needs to perform about 6N floating point multiplications in transferring the Cartesian coordinates to lattice coordinates, where N is the number of atoms. With this method we can calculate the SASA of N-terminal domain of the 434 repressor (1R69, 484 nonhydrogen atoms) to within 1.2% accuracy in an average CPU time of 0.115 s on a SGI PowerChallege R10000 processor at 194Mhz.

Conformational Space Searching Algorithm and Supersecondary Structure Constraints

We use genetic algorithm (GA)[42,43] to search the peptide chain conformational space for low-energy structures. In recent years there have been many

TABLE II. Corresponding Regions of the Supersecondary Structure Constraints†

Supersecondary structures	ϕ	ψ
α-helix	$-75° \sim -55°$	$-50° \sim -30°$
β-strand	$-130° \sim -110°$	$110° \sim 130°$
a	$-150° \sim -30°$	$-100° \sim 50°$
b	$-230° \sim -30°$	$100° \sim 200°$
e	$30° \sim 130°$	$130° \sim 260°$
l	$30° \sim 150°$	$-60° \sim 90°$
t	$-160° \sim -50°$	$50° \sim 100°$
undefined*	$-180° \sim 0°$**	$-180° \sim 180°$

†The rectangles were used as substitutes for the irregular regions of a, b, e, l, and t on the ϕ-ψ map. The most populated area of each region was included in the corresponding rectangle.
*The residues that did not belong to any predicted supersecondary structures were classified as undefined.
**For glycine this should be $-180° \sim 180°$.

studies of the use of GAs for protein structure prediction and other related structure optimization problems.[12,44-49] The basic idea of the genetic algorithm is to give better chances of survival and reproduction to the good individuals of the population (in our case, the low-energy structures). In this way the good genes (structural factors) will accumulate and combine gradually to dominate the whole population. In the GA used in this work, a chromosome consists of all the free variables in our peptide chain representation, i.e., it encodes the set of ϕ, ψ, and χ. Most residues have one, two, or three sidechain torsion angles, so the length of the chromosome is about 4N, where N is the number of residues. There are many versions of GA in its applications to protein folding simulation. Our GA procedure is:

The initial population

The initial population size was 500. These structures were built by randomly selecting the backbone and sidechain torsion angles in the constrained regions. The backbone torsion angles ϕ and ψ of a residue were sampled uniformly in a certain region with a fineness of 1°. This region was defined by the position of the residue in the predicted supersecondary structures (Table II). The backbone torsion angles of the residues in the short connecting peptide of any super-secondary structure were constrained to lie in the corresponding regions on the ϕ-ψ map (Fig. 1). The backbone dihedral angles of the residues in the α-helix and β-strand were restrained to within ±10° of their ideal value, which was set to (−65°, −40°) and (−120°, 120°) respectively. In this way, predicted supersecondary structures were used to greatly reduce the allowable variation of the affected torsion angles. For the other residues not predicted to lie in any supersecondary structure, their backbone torsion angles are sampled randomly on the left half ($\phi < 0$) of the ϕ-ψ map (for glycine, it was the whole ϕ-ψ map). For the sidechain torsion angles, the constraints were based on the sidechain rotamer library.[41] The mean value of the sidechain torsion angle in the rotamer library was selected according to its occurring ratio. After a mean value was selected, an integer value is selected randomly from the interval [mean value − standard deviation, mean value + standard deviation]. These 500 structures were the parent individuals of the first generation.

Fitness criterion

The potential energy of the 500 parent individuals were computed. Potential energy was used as the objective function. Then we mapped the objective function onto a fitness scale. If there are a few extraordinary individuals in the early stage of the GA process, they will take over a significant proportion of the population after several generations. This is a leading cause of premature convergence. On the other hand, if there is still significant diversity within the population in the later stages of the GA process, then the population average fitness may be close to the population best fitness and the best members may not become dominant in the population. In this case, the GA process becomes a random walk.[43] To prevent premature convergence and random walk, we used a generation-dependent fitness scaling.

$$Fitness_{gn,i} = 1 + C_{gn}\frac{E_{gn,\max} - E_{gn,i}}{E_{gn,\max} - E_{gn,\min}}$$

$$C_{gn} = C_0 + incr \cdot gn$$

where $Fitness_{gn,i}$ is the fitness of the *ith* individual in *gnth* generation, $E_{gn,max}$, $E_{gn,min}$, and $E_{gn,i}$ is highest, lowest, and the *ith* individual's potential energy in the *gnth* generation, C_0 is a constant that is set to be 0.02, *incr* is increment of the ratio of fitness of the best individual (with lowest energy) to the worst individual (with highest energy in each generation). We set $incr = 0.0016$ in our computations.

In each generation, the fitness of the worst individual was always set to 1, the best individual was $1 + C_0 + incr \cdot gn$. This scaling strategy is a variant of the fitness scaling methods that focus on the ratio of the fitness of the best individual and the average fitness. Premature convergence and random walk can be prevented by this scaling strategy.

The crossover operation

Pairs of individuals were selected randomly for crossover operation. The probability for an individual to be selected was $f_i/\Sigma f$, where f_i was the fitness of the individual and Σf was the summation of the fitness of all the individuals in the population.

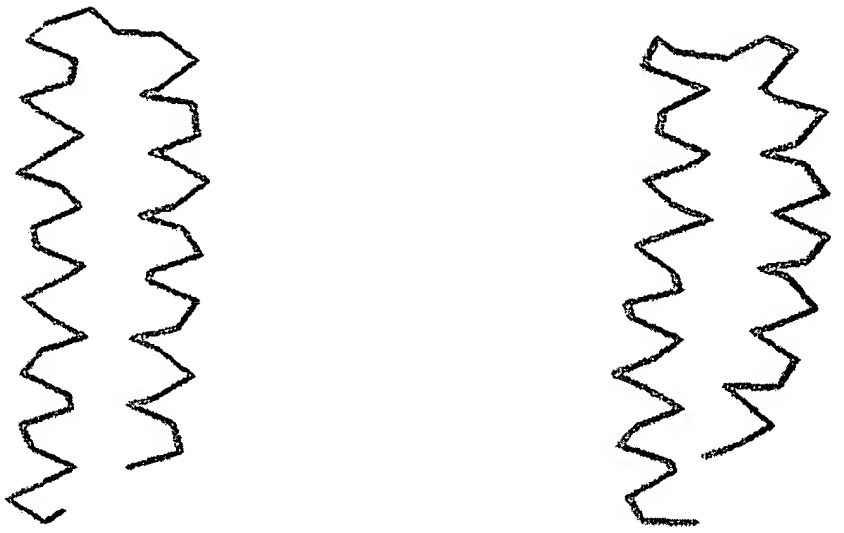

Fig. 3. Structure comparison between the crystal structure (native) and the computed structure (predicted) of repressor of primer.

The chromosome looked like:

$$\phi_1\psi_1\chi_{11}\chi_{12}|\phi_2\psi_2\chi_{21}\chi_{22}\chi_{23}|\ \phi_3\psi_3\chi_{31}|$$
$$\cdots\cdots|\ \phi_n\psi_n\chi_{n1}\chi_{n2}|\ \cdots\cdots|\phi_N\psi_N\chi_{N1}\chi_{N2},$$

where n was the residue number, χ_{n1} and χ_{n2} was the χ_1 and χ_2 of the *nth* residues, | is a possible crossover site. In order to keep the correlation between ϕ and ψ in the super-secondary structures, and the correlation of χ_1, χ_2, and χ_3 in the sidechain rotamer library, the crossover site is not allowed to occur between the ϕ, ψ, and χ of the same residue. A selected pair of chromosomes would undergo a fixed number (chosen to be 1 in this study) of crossovers at randomly chosen allowable sites.

The mutation operation

Two kinds of mutation operators were used. The first mutation operator may change the conformation dramatically. When this operator acted on a peptide chain, all the values of the backbone and sidechain torsion angles of a randomly chosen residue were reselected from their corresponding constrained regions. We made a copy of the 500 parent individuals and modified this copied population M_1 times, each time by applying the operator to a randomly selected individual from this population. An individual can be selected more than one time, so there may be changes in torsion angles in more than one residue in a chromosome. The second mutation operator is for a more local search of conformational space.[12] It will perturb some residues' torsion angles (ϕ, ψ, and χ) by a random angle between $-5°$ and $5°$. The number of perturbed residues of each individual is M_2. This operator was also applied to every parent individual so that in total 500 offspring were produced.

Again, for the purposes of preventing premature convergence and random walk, we made M_1 and M_2 decrease as the search proceeds:

$$M_1 = 1 + P \cdot \exp(-gn/Neff)$$

$$M_2 = 1 + \frac{N}{4} \cdot \exp(-gn/Neff)$$

where P is set to 500, N is the number of residues in the protein, gn is the generation, and $Neff$ is a constant set to 150.

Selection

Now the population consists of 500 parent conformations, 500 crossed offspring, and 1,000 mutated offspring. The total population is 2,000. The potential energy of these 2,000 conformations were computed and only the 500 lowest-energy conformations were selected into the next generation as parent conformations.

Convergence

At least 100 generations of GA were performed for each protein. After 100 generations, the GA process will stop only if the decrease of the lowest energy in the population is less than 1 unit during the last 20 generations. On average, about 150 generations of GA were performed for each protein.

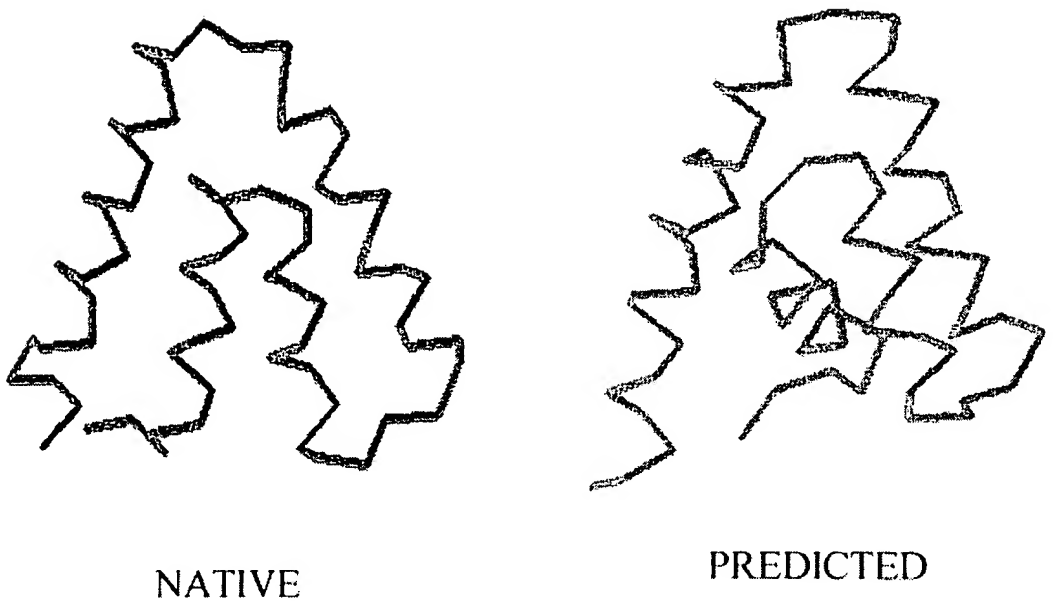

Fig. 4. Structure comparison between the crystal structure (native) and the computed structure (predicted) of Uteroglobin.

RESULTS

Repressor of Primer (1ROP)

Repressor of primer is a 4-helix bundle protein that is composed of two identical monomers. Each monomer has 56 residues and forms a α–turn–α structure, which does not belong to the 11 frequently occurring supersecondary structures. The predicted secondary structures (Fig. 2a) were used as constraints. After computing the conformation using our algorithm, we calculated its distance matrix error (DME) to the crystal structure. The computed structure matches the crystal structure with a DME of 1.48 Å (Fig. 3).

Uteroglobin (1UTG)

Uteroglobin is a 4-helix protein that has 70 residues. The predicted supersecondary structures are α-bb-α-lbb-α-bb-α (Fig. 2b). With these supersecondary structure constraints we computed the 3-dimensional structure. The computed structure matches the crystal structure with a DME of 3.47 Å (Fig. 4).

Crambin (1CRN)

Crambin is a 46-residue protein with two α-helix and a pair of β-strands. It has three disulphide bonds. We did not use the disulphide bond constraints. The predicted supersecondary structures are β-loop-α-lbb-α-l-β-loop-α (Fig. 2c). The computed structure matches the crystal structure with a DME of 2.73 Å (Fig. 5).

N-Terminal Domain of the 434 Repressor (1R69)

The crystal structure of the N-terminal domain of the 434 repressor has 63 residues and is composed of five helices. The predicted supersecondary structures are α-lbb-α-lbb-α-loop-α-lbb-α (Fig. 2d). The computed structure matches the crystal structure with a DME of 4.48 Å (Fig. 6).

C-Terminal Domain of the L7(SLASH)*L12 50 S Ribosomal Protein (1CTF)

This protein has 68 residues. It has six secondary structures—three α-helix and three β-strands. This protein is the most complex example in this study. The predicted supersecondary structures are β-α-lbb-α-β-α-l-β (Fig. 2e). The computed structure matches the crystal structure with a DME of 4.00 Å (Fig. 7).

DISCUSSION

The predicted super-secondary structures and the native supersecondary structures of these five proteins are shown in Fig. 2. In these five proteins, there are 21 secondary structures and 16 short connecting peptides. Ten short connecting peptides were identified to be in one of the 11 frequently occurring supersecondary structures. Most of the supersecondary structures are correctly predicted. For these five proteins the correctness ratio is 90.1%. Although the accuracy is high, in some instances the predicted structures do not align precisely with those observed in the crystal structures. If the backbone torsion angle (ϕ, ψ) of a few consecutive residues were

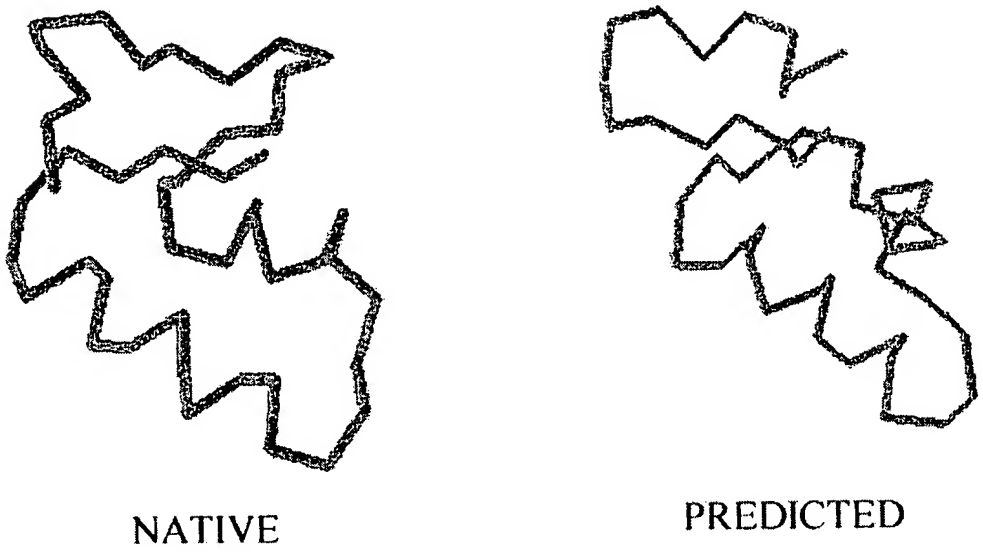

Fig. 5. Structure comparison between the crystal structure (native) and the computed structure (predicted) of repressor of Crambin.

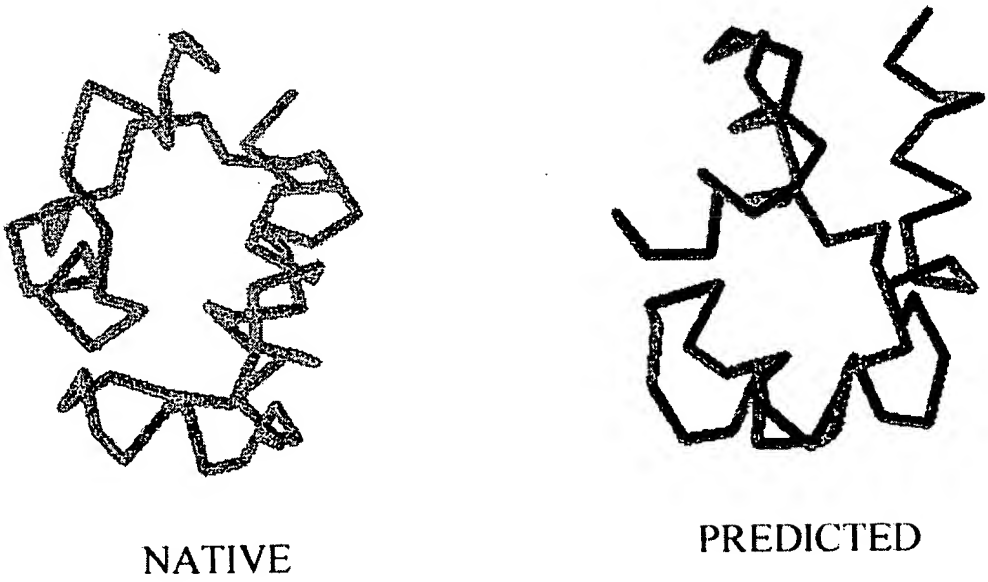

Fig. 6. Structure comparison between the crystal structure (native) and the computed structure (predicted) of N-terminal domain of the 434 repressor.

restrained in wrong regions, the peptide chain may have a wrong trend at this segment. If the segment was at the central part of the peptide chain, the overall fold may be misdetermined. In this study, such a fatal mistake has not occurred in the supersecondary structure prediction. The most serious structure distortion caused by the errors of the supersecondary structure prediction was in the crambin. At the C-terminal of the peptide chain, an incorrectly predicted α-helix (from residue 41 to 45) was imposed on the peptide chain as a constraint (Fig. 4). As a result, a wrong structure was formed in this terminal (Fig. 7).

In Figure 2 one can find that in some cases supersecondary structure was not correctly predicted at only one or two residues, while the neighboring residues were all restrained in the correct regions. In these cases, the residue the peptide chain will turn to a wrong direction at this point. But if the native-like structures are favored by the potential, the nearby residues will move to compensate for this mistake. As a result, a native-like profile can still be

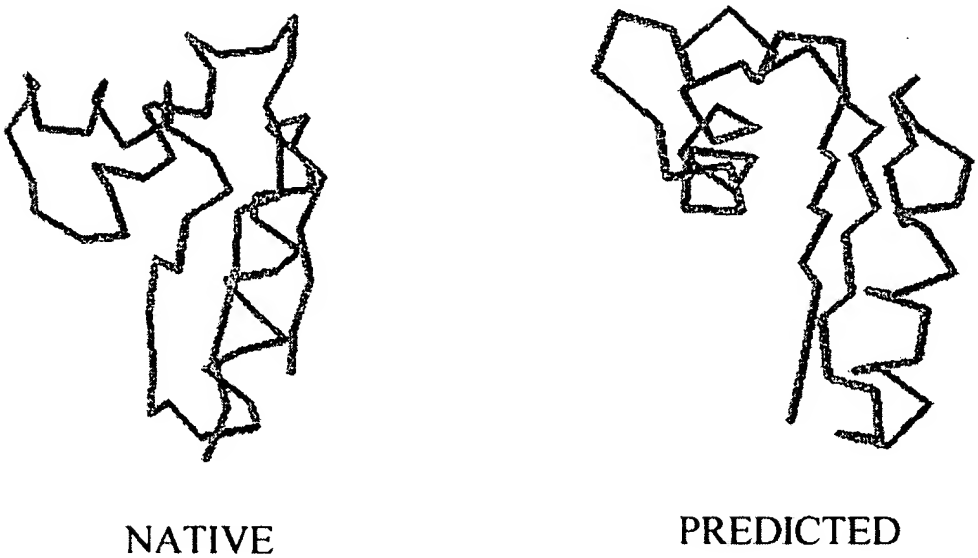

Fig. 7. Structure comparison between the crystal structure (native) and the computed structure (predicted) of C-terminal domain of the ribosomal protein L7/*L12.

formed. For example, the computed structure of lutg (Fig. 4) was native-like, while at two central residues (15 and 28) the supersecondary structures were predicted incorrectly. This is possible because of the flexibility of the model. In this model, the peptide chains are more flexible than those in the fixed secondary structure model[12] where the affected torsion angles are fixed at their ideal values (without any degree of freedom). In contrast, the torsion angles in our model can rotate in some regions, which is determined by the predicted supersecondary structures and the sidechain rotamer library.

With the supersecondary structure constraints the conformation space of the peptide chain is greatly reduced, but the DME of the peptide chain can still be large (over 15 Å). For example, in the simulation of the 434 repressor we observed many conformations with a DME exceeding 15 Å, especially in the early generations. This indicates that the overall fold of a protein molecule cannot be well defined only by the supersecondary structures. Genetic algorithm was used to search the reduced conformation space for low-energy structures. Four of the five computed structures are similar to the corresponding X-ray elucidated structures. The DME of repressor of primer, uteroglobin, crambin, and C-terminal domain of L7(SLASH)*L12 50 S ribosomal protein are all smaller than or equal to 4.0 Å. For the N-terminal domain of the 434 repressor, there are three supersecondary structures. One of the connecting peptides, a six-residue loop, cannot be recognized as belonging to any of the 11 kinds of supersecondary structures. The peptide chain is divided into two fragments which are connected by the loop region. The computed structure of each of these two fragments is similar to their corresponding parts in the crystal structure, but the relative position of the two fragments is incorrectly determined by the loop region.

An ideal potential should give higher values to non-native conformations than the native conformation. Recent studies[50-52] indicated that some potentials can distinguish between correct and certain incorrect structures with a high degree of success. However, in the protein folding simulation there is an astronomical number of candidates in the conformation space. If a potential can identify 99.99% non-native structures, while it gives 0.01% of them lower energy than the native structure, then there are still uncountably many minima with lower energy than that of the native structure on the energy landscape. In this situation, our hope of finding a native-like structure lies in the possibility that most of the low-energy structures are "near" the native structure to form a cluster of native-like structures. This appears to be the case for the repressor of primer and uteroglobin, where the energy of the computed structures is lower than that of their native structures. The other three proteins are more complex; the energy of the computed structures are higher than that of their native structures. This is mainly caused by the van der Waals term and it indicates that a more efficient local search method is needed.

In recent years, important progress has been made in computing the 3-dimensional structures of proteins from their sequences using simple energy function. The attractive aspect of this method is that a simple model should be much easier to improve than highly parameterized ones, and the prediction

results of these simple models is arguably comparable to those of the more complex models.[14] Encouraging results have been reported by Sun et al.[12] They developed a model that predicted reasonably well the known tertiary folds of 7 out of 10 small proteins. Their method used experimental secondary structures, in which the backbone dihedral angles (ϕ, ψ) are fixed at the ideal values. Three of these ten proteins are also considered in this study. They are repressor of primer, crambin, and N-terminal domain of the 434 repressor. The DME reported in Sun et al. are 1.65 Å, 4.87 Å, and 5.55 Å, respectively. In our results, the DME of these proteins are 1.48 Å, 2.73 Å, and 4.48 Å, respectively. This suggests that supersecondary structure constraints and better modeling of the hydrophobic interaction are of considerable utility in protein structure computation.

CONCLUSION

One important step toward building a tertiary structure is to identify how secondary structures as building blocks arrange themselves in space. Good supersecondary structure prediction methods can provide important information in the prediction of protein tertiary structure.

The structure of a protein is determined by the competition and cooperation of all of the interactions, especially hydrophobic interaction and van der Waals interaction. Correctly including the hydrophobic interaction is extremely important. Although the nature of hydrophobic interaction is not completely understood, it is suggested that protein–solvent interaction depends on the solvent-accessible surface area of the protein molecule.[53,54]

The goal of this study is to suggest a way to capture these two main features in our current understanding of protein structure, interaction, and folding mechanism. The results show that some small protein structures can be determined by a model that carefully adduces these points.

ACKNOWLEDGMENTS

We thank the National Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics & Scientific and Engineering Computing, Chinese Academy of Sciences, and the Computer Network Information Center, Chinese Academy of Sciences, for providing free CPU time. We are grateful to Professor Zhirong Sun for his help in the prediction of the supersecondary structures of the five proteins used in this study.

REFERENCES

1. Richards, F.M. Areas, volumes, packing, and protein structures. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
2. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 14:1–64, 1959.
3. Dill, K.A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.
4. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S. Principle of protein folding—A perspective from simple exact models. Protein Sci. 4:561–602, 1995.
5. Eisenberg, D., Weiss, R.M., Terwillinger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. USA 81:140–144, 1984.
6. Crippen, G.M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
7. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
8. Wetlaufer, D. Nucleation, rapid folding, and globular interchain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
9. Levinthal, C. Are there pathways in protein folding? J. Chem. Phys. 65:44–45, 1968.
10. Unger, R., Moult, J. Finding the lowest free energy conformation of a protein is a NP-hard problem: Proof and implication. Bull. Math. Biol. 55:1183–1198, 1993.
11. Sun, Z., Rao, X., Peng, L., Xu, D. Prediction of protein supersecondary structures based on the artificial neural network method. Protein Eng. 10:763–769, 1997.
12. Sun, S., Thomas, P.D., Dill, K.A. A simple protein folding algorithm using a binary code and secondary structure constraints. Protein Eng. 8:769–778, 1995.
13. Srinivasan, R., Rose, G. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.
14. Yue, K., Dill, K.A. Folding proteins with a simple energy function and extensive conformational searching. Protein. Sci. 5:254–261, 1996,
15. Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S., Blundell, T.L. Fragment ranking in modelling of protein structure: Conformationally constrained environmental amino acid substitution tables. J. Mol. Biol. 229:194–220, 1993.
16. Sun, Z., Jiang, B.J. Patterns and conformations commonly occurring supersecondary structures (basic motifs) in Protein Data Bank. J. Protein Chem. 15:675–690, 1996.
17. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
18. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
19. Richarmond, T.J., Richards, F.M. Packing of α-helices: Geometrical constraints and contact areas. J. Mol. Biol. 119:537–555, 1978.
20. Finney, J.L. Volume occupation, environment, and accessiblity in proteins: Environment and molecular area of RNase-S. J. Mol. Biol. 119:415–441, 1978.
21. Greer, J., Bush, B. Macromolecular shape and surface maps by solvent exclusion. Proc. Natl. Acad. Sci. USA 75:303–307, 1978.
22. Pearl, L.H., Honegger, A. Generation of molecular surfaces for graphic display. J. Mol. Graph. 1:9–12, 1983.
23. Mueller, J.J. Calculation of scattering curves for macromolecules in solution and comparison with results of methods using effective atomic scattering factors. J. Appl. Cryst. 16:74–82, 1983.
24. Pavlov, M. Y., Fedorov, B.A. Improved technique for calculating X-ray scattering intensities in solution: Evaluation of the form, volume, and surface of a particle. Biopolymers 22:1507–1522, 1983.
25. Lorensen, W., Cline, H. Marching cubes: A high resolution 3D surface construction algorithm. Comput. Graph. 21:163–169, 1987.
26. Meyer, A.Y. Molecular mechanics and molecular shape. V. On the computation of the bare surface area of molecules. J. Comp. Chem. 9:18–24, 1988.
27. Karfunkel, H.R., Eyrand, V. An algorithm for the representation and computation of supermolecular surfaces and volumes. J. Comp. Chem. 10:628–634, 1989.
28. Connolly, M.L. Analytical molecular surface calculation. J. Appl. Cryst. 16:548–558, 1983.
29. Richmond, T.J. Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. J. Mol. Biol. 178:63–89, 1984.

30. Connolly, M.L. Molecular surface triangulation. J. Appl. Cryst. 18:499–505, 1985.
31. Gibson, K.D., Scheraga, H.A. Exact calculation of the volume and surface area of fused hard-sphere molecules with unequal atomic radii. Mol. Phys. 62:1247–1265, 1987.
32. Gibson, K.D., Scheraga, H.A. Surface area of the intersection of three sphere with unequal radii: a simplified analytical formula. Mol. Phys. 64:641–644, 1988.
33. Dodd, L.R., Theodorou, D.N. Analytical treatment of the volume and surface area of molecules formed by an arbitrary collection of unequal spheres intersected by planes. Mol. Phys. 72:1313–1345, 1991.
34. Wang, H., Levinthal, C. A vectorized algorithm for calculating the accessible surface area of macromolecules. J. Comp. Chem. 12:868–871, 1991.
35. Pascual-Ahuir, J.L., Silla, E. GEPOL: An improved description of molecular surfaces. I. Building the spherical surface set. J. Comp. Chem. 11:1047–1060, 1991.
36. Silla, E.J., Tunon, I., Pascual-Ahuir, J.L. GEPOL: An improved description of molecular surfaces. II. Computing the molecular area and volume. J. Comp. Chem. 12:1077–1088, 1991.
37. Perrot, G., Cheng, B., Gibson, K.D., et al. MSEED: A program for the rapid analytical determination of accessible surface areas and their derivatives. J. Comp. Chem. 13:1–11, 1992.
38. LeGrand, S.M., Merz, K.M.M. Jr., Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. J. Comp. Chem. 14:349–352, 1993.
39. Eisenhaber, F., Argos, P., Sander, C., Scharf, C. The double cubic lattice method: Efficent approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comp. Chem. 16: 273–284, 1995.
40. Totrov, M. The contour-buildup algorithm to calculate the analytical molecular surface. J. Struct. Biol. 116:138–143, 1996.
41. Ponder, J.W., Richards, F.M. Tertiary templates for proteins use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
42. Holland, J. "Adaptation in Natural and Artificial Systems." Ann Arbor, MI: University of Michigan Press, 1975.
43. Goldberg, D.E. "Genetic Algorithm in Search, Optimization and Machine Learning." Reading, MA: Addison-Wesley, 1989.
44. Unger, R., Moult, J. Genetic algorithm for protein folding simulation. J. Mol. Biol. 231:75–81, 1993.
45. Sun, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci. 2:762–785, 1993.
46. Bowie, J.U., Eisenberg, D. An evolutionary approach to folding proteins from sequence information: Application to small α-helical proteins. Proc. Natl. Acad. Sci. USA 91:4436–4440, 1994.
47. Dandekar, T., Argos, P. Folding the main chain of small proteins with the genetic algorithm. J. Mol. Biol. 236:844–861, 1994.
48. Pedersen, J.T., Moult, J. Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. Proteins 23:454–460, 1995.
49. Pedersen, J.T., Moult, J. Protein folding simulations with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269:240–259, 1997.
50. Wang, Y., Zhang, H., Li, W., Scott, R.A. Discriminating compact nonnative structures from the native structure of globular proteins. Proc. Natl. Acad. Sci. USA 92:709–713, 1995.
51. Huang, E.S., Subbish, S., Tsai, J., Levitt, M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J. Mol. Biol. 257:716–725, 1996.
52. Park, B., Levitt, M. Energy function that discriminate X-ray and near native folds from well-constructed decoys. J. Mol. Biol. 258:367–392, 1996.
53. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. Nature 248:338–339, 1974.
54. Eisenberg, D., McLanchlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.